Using Thesauri in Cross-Language Retrieval of German and French Indexed Collections

Vivien Petras[1], Natalia Perelman[1], and Fredric C. Gey[2]

[1]*School of Information Management and Systems*
[2]*UC Data Archive & Technical Assistance*
University of California, Berkeley, CA 94720 USA

This is an author's accepted manuscript version of a conference paper published in *International Conference of the Cross-Language Evaluation Forum for European Languages, CLEF 2002: Advances in Cross-Language Information Retrieval* within the Springer Lecture Notes in Computer Science book series (LNCS, volume 2785).

The final publisher's version is available online at:
https://doi.org/10.1007/978-3-540-45237-9_31

Using Thesauri in Cross-Language Retrieval of German and French Indexed Collections

Vivien Petras[1] Natalia Perelman[1] and Fredric Gey[2]

[1] School of Information Management and Systems
[2] UC Data Archive & Technical Assistance
University of California, Berkeley, CA 94720 USA

Abstract. For CLEF-2002, Berkeley's group one experimented with Russian, French and English as query languages, and investigated thesaurus-aided retrieval for the special CLEF collections GIRT and Amaryllis. Two techniques were used to locate source language topic terms within the controlled vocabulary and replace them with the document language thesaurus terms to form the query sent against the collection index. This form of controlled vocabulary-aided translation is called thesaurus matching. Results show that thesaurus-aided cross-language retrieval performs slightly worse than machine translation retrieval on average, but can yield decidedly better results for particular queries.
In addition, Berkeley submitted runs to the monolingual and bilingual (French and German) CLEF main tasks. We found that bilingual retrieval sometimes outperforms monolingual retrieval and postulate reasons to explain this phenomenon.

1 Introduction

Digital libraries relating to particular subject domains have invested a great deal of human effort in developing metadata in the form of subject area thesauri. This effort has emerged more recently in artificial intelligence as ontologies or knowledge bases which organize particular subject areas. The purpose of subject area thesauri is to provide organization of the subject into logical, semantic divisions as well as to index document collections for effective browsing and retrieval. Prior to free-text indexing (i.e. the bag-of-words approach to information retrieval), subject area thesauri provided the only point of entry (or 'entry vocabulary') to retrieve documents. A debate began over thirty years ago about the relative utility of the two approaches to retrieval:

- to use index terms assigned by a human indexer, drawn from the controlled-vocabulary, or
- to use automatic free-text indexing from the words or phrases contained in the document text.

This debate continues to this day and the evidence seems to have been mixed. In performance studies thesaurus-aided retrieval performs worse than free-text over a group of queries, while it performs better for particular queries [1].

It is an interesting question to evaluate what utility and performance can be obtained in cross-language information retrieval (CLIR) with the use of multilingual thesauri. The two domain-specific CLEF tasks, Amaryllis and GIRT, provide the opportunity to examine CLIR performance for such thesauri. The GIRT task provides a thesaurus for the social sciences in German, English, and (by translation) Russian, and Berkeley has studied it for three years. Amaryllis does not have a thesaurus per se (i.e. it does not identify broader terms, narrower terms or related terms), but it does have a specialized controlled vocabulary for its domain of coverage in both the French and English languages.

We have been evaluating thesaurus-aided retrieval by comparing traditional machine translation with our thesaurus matching techniques. We match a non-source-collection language search topic against the non-source-collection language version of our thesaurus (e.g. look up English query words in the English version of the GIRT thesaurus). Once a match is found, we replace the non-source-collection language thesaurus term with the source-collection language thesaurus term (e.g. the English GIRT thesaurus term is replaced with the associated German GIRT thesaurus term) and search those against the collection.

In addition we have been investigating the viability of Russian as a query language for the CLEF collections and continue this research for the CLEF bilingual (Russian to German and Russian to French) main tasks and the GIRT task (Russian to German).

For monolingual retrieval the Berkeley group has used the technique of logistic regression from the beginning of the TREC series of conferences. In the TREC-2 conference [2] we derived a statistical formula for predicting probability of relevance based upon statistical clues contained within documents, queries and collections as a whole.

2 Amaryllis

The Amaryllis task consisted of retrieving documents from the Amaryllis collection of approximately 150,000 French documents which were abstracts of articles in a broad range of disciplines (e.g. biological sciences, chemical sciences, engineering sciences, humanities and social sciences, information science, medical sciences, physical and mathematical sciences, etc). There were twenty-five topics and the primary goal was French-French monolingual retrieval under multiple conditions (primarily testing retrieval with or without concept words from the Amaryllis controlled vocabulary). An auxiliary task was to test out English to French cross-language information retrieval.

The first French topic is found below as Figure 1. Note that in distinction from previous CLEF tasks, the narrative field (FR-narr) consists only of controlled vocabulary concepts taken from the Amaryllis controlled vocabulary rather than the usual narration which expands upon the description.

Fig. 1. Amaryllis Topic 01 (French and English xml)

```
<top>
  <num>001</num>
  <FR-title>Impact sur l'environnement des moteurs diesel</FR-title>
  <FR-desc>Pollution de l'air par des gaz d'échappement des moteurs diesel
    et méthodes de lutte antipollution. Emissions polluantes (NOX, SO2,
    CO, CO2, imbrûlés, ...) et méthodes de lutte antipollution</FR-desc>
- <FR-narr>
    <c>Concentration et toxicité des polluants</c>
    <c>Mécanisme de formation des polluants</c>
    <c>Réduction de la pollution</c>
    <c>Choix du carburant</c>
    <c>Réglage de la combustion</c>
    <c>Traitement des gaz d'échappement</c>
    <c>Législation et réglementation</c>
  </FR-narr>
</top>
<top>
  <num>001</num>
  <EN-title>The impact of diesel engine on environment</EN-title>
  <EN-desc>Air pollution by the exhaust of gas from diesel engines and
    methods of controlling air pollution. Pollutant emissions (NOX, SO2, CO,
    CO2, unburned product, ...) and air pollution control</EN-desc>
- <EN-narr>
    <c>Concentration and toxicity of pollutant</c>
    <c>Pollutant formation mechanism</c>
    <c>Pollution prevention and reduction</c>
    <c>Motor fuel selection</c>
    <c>Combustion control</c>
    <c>Exhaust gas treatment</c>
    <c>Legislation and regulation</c>
  </EN-narr>
</top>
```

For the Amaryllis task, we experimented with the effects of translation, inclusion of concept words and thesaurus matching. We indexed all fields in the document collection and used a stop word list, the latin-to-lower normalizer (changes capitals into lower case) and the Muscat French stemmer.

2.1 Amaryllis Controlled Vocabulary Matching

For Amaryllis controlled vocabulary matching we first extracted individual words and phrases from the English topics. Phrases were identified by finding the longest matching word sequences in the Amaryllis vocabulary file that was used as a segmentation dictionary. This method identified phrases such as "air pollution" and "diesel engine" in the first topic. The individual words and phrases were then searched in the Amaryllis vocabulary and if a match was found the words were replaced with their French equivalents.

2.2 Amaryllis Runs

Our Amaryllis results are summarized in Table 1. The runs are described below. The performance is computed over the top ranked 1000 documents for 25 queries.

Table 1. Results of official Amaryllis runs for CLEF-2002.

Run Name	BKAMFF1	BKAMFF2	BKAMEF1	BKAMEF2	BKAMEF3
Qry ndx	TDN	TD	TD	TD	TD
Retrieved	25000	25000	25000	25000	25000
Relevant	2018	2018	2018	2018	2018
Rel Ret	1935	1863	1583	1897	1729
Precision					
at 0.00	0.9242	0.8175	0.6665	0.8079	0.6806
at 0.10	0.8011	0.7284	0.5198	0.7027	0.6497
at 0.20	0.7300	0.6296	0.4370	0.6114	0.5874
at 0.30	0.6802	0.5677	0.3791	0.5612	0.5337
at 0.40	0.6089	0.5159	0.3346	0.5033	0.4983
at 0.50	0.5458	0.4722	0.2942	0.4489	0.4452
at 0.60	0.4784	0.4035	0.2481	0.3825	0.3848
at 0.70	0.4242	0.3315	0.1874	0.3381	0.3114
at 0.80	0.3326	0.2682	0.1251	0.2664	0.2414
at 0.90	0.2193	0.1788	0.0501	0.1888	0.1570
at 1.00	0.0596	0.0396	0.0074	0.0300	0.0504
Avg Prec.	0.5218	0.4396	0.2792	0.4272	0.4038

BKAMFF1, our monolingual run including the concepts in the queries (title, description and narrative) yielded the best results. Our second monolingual run, BKAMFF2, where we excluded the concepts from the query indexes (only title and description) resulted in a 20% drop in average precision. Blind feedback improved the performance for both runs.

In comparing thesaurus matching and translation, this year the translation runs yielded better results. As a baseline, we run the English Amaryllis queries (without concepts or translation) against the French Amaryllis collection (BKAMEF1). As expected, average precision was not very high, but it is still greater than 50 percent of the best monolingual run. Using machine translation for the second bilingual run (BKAMEF2) improved precision over 50%. For translating the English topics, we used the Systran and L & H Power translator. By using only the Amaryllis thesaurus to match English words with French thesaurus terms (the BKAMEF3 run), we improved our average precision 44% compared to the baseline. For all runs, the query indexes only included the title and description fields, but we used blind feedback for BKAMEF2 and BKAMEF3.

3 GIRT task and retrieval

The GIRT collection consists of reports and papers (grey literature) in the social science domain. The collection is managed and indexed by the GESIS organization (http://www.social-science-gesis.de). GIRT is an excellent example of a collection indexed by a multilingual thesaurus, originally German-English, recently translated into Russian. The GIRT multilingual thesaurus (German-English), which is based on the Thesaurus for the Social Sciences [3], provides the vocabulary source for the indexing terms within the GIRT collection of CLEF. There are 76,128 German documents in the GIRT subtask collection. Almost all the documents contain manually assigned thesaurus terms. On average, there are about 10 thesaurus terms assigned to each document. More detail about the GIRT task may be found in [4].

For the Girt task, we experimented with the effects of different thesaurus matching techniques and the inclusion of thesaurus terms. The German Girt collection was indexed using the German decompounding algorithm to split compounds (see section 4). For all runs, we used our blind feedback algorithm to improve the runs' performance.

3.1 GIRT Thesaurus Matching

Similar to the Amaryllis thesaurus-based translation, we initially identified some phrases in the English GIRT topics by finding the longest matching entries in the English-German GIRT thesaurus. This method produced phrases such as "right wing extremism" and "drug abuse". Then individual topic words and phrases were matched against the thesaurus and replaced with their German translations.

For the thesaurus-based translation of Russian GIRT topics we first transliterated both Russian topics and Russian entries in the German-Russian GIRT thesaurus by replacing Cyrillic characters with their Roman alphabet equivalents. Then two different approaches were used to find matches in the thesaurus.

In the first approach (run BKGRRG1 below) we identified phrases by finding the longest sequences of exact word matches to the terms in the thesaurus. The second approach was to follow the exact match with a fuzzy matching method to match both phrases and individual words that were not identified by the exact match match method. This method, previously employed by our group in CLEF 2001, identified thesaurus terms in Russian by determining Dice's coefficient of similarity between the topic words and phrases and the thesaurus entries [5]. Since fuzzy matching sometimes finds commonality between unrelated words, in our second approach, in order to deal with Russian inflectional morphology, we normalized Russian words by removing the most common Russian inflectional suffixes. Then we identified phrases as in the previous method and translated both phrases and individual words by finding their matches in the thesaurus.

The following examples delineate the differences between the fuzzy match and longest match:

Russian word	Fuzzy Thesaurus Match	Exact match	German translation
uverennost'	samouverennost'	(no match)	Selbstsicherheit
devushek	devushka	(no match)	weibliche Jugendliche
uchebnye	uchebnik	ucheba	Lehrbuch–Lernen
sotsial'noi	sotsial'nyi sloi	(no match)	soziale Schicht
polov	(no match)	pol	Geschlecht
shkol'nikov	doshkol'nik	(no match)	Vorschulkind
obshchestvennye	obshchestvennye nauki	(no match)	Gesellschaftswissenschaft

While fuzzy matching finds more matches, it sometimes overgenerates: for example, it matched the original topic word "shkol'nik" (schoolboy) with "doshkol'nik" (preschool aged child).

3.2 GIRT results and analysis

Our GIRT results are summarized in Table 2. We had five runs, two monolingual and three cross-language, two with Russian topics, and one with English topics. The runs are described below. Only 24 of the 25 GIRT queries had relevant documents, consequently the performance is computed over the top ranked 1000 documents for 24 queries. Except for the second monolingual run (BKGRGG2),

Table 2. Results of official GIRT runs for CLEF-2002.

Run Name	BKGRGG1	BKGRGG2	BKGREG1	BKGRRG1	BKGRRG2
Qry ndx	TDN	TD	TD	TDN	TD
Retrieved	24000	24000	24000	24000	24000
Relevant	961	961	961	961	961
Rel. Ret	853	665	735	710	719
Precision					
at 0.00	0.7450	0.6227	0.5257	0.5617	0.5179
at 0.10	0.6316	0.4928	0.3888	0.3595	0.3603
at 0.20	0.5529	0.4554	0.3544	0.3200	0.3233
at 0.30	0.5112	0.3551	0.3258	0.2705	0.2867
at 0.40	0.4569	0.3095	0.2907	0.2275	0.2263
at 0.50	0.4034	0.2462	0.2345	0.1793	0.1932
at 0.60	0.3249	0.1953	0.2042	0.1451	0.1553
at 0.70	0.2753	0.1663	0.1432	0.0945	0.1105
at 0.80	0.2129	0.1323	0.1188	0.0679	0.0858
at 0.90	0.1293	0.0497	0.0713	0.0413	0.0606
at 1.00	0.0826	0.0216	0.0454	0.0256	0.0310
Avg Prec.	0.3771	0.2587	0.2330	0.1903	0.1973

we indexed all allowed fields (including the controlled terms) in the document collection.

Using all query fields and indexing the controlled terms resulted in a 45% improvement in average precision for the monolingual Girt run BKGRGG1 compared to BKGRGG2 (which only indexed the title and description query fields).

For Russian-German retrieval, the positive effect of including the narrative fields for the query indexing was countered by the different thesaurus matching techniques for the Russian Girt run. Although the BKGRRG1 run used all query fields for searching, its fuzzy thesaurus matching technique resulted in a 3% drop in average precision compared to the BKGRRG2 run, which only used the title and description topic fields for searching but used a different thesaurus matching technique. Both runs pooled 2 query translations (Systran and Promt) and the thesaurus matching results into one file.

Comparing the Russian Girt runs (translation plus thesaurus matching) to the Russian to German bilingual runs with translation only (also: different collection), one can see a 36% and 70% improvement in average precision for the title and description only and the title, description and narrative runs, respectively.

Our final Girt run was BKGREG1 (Berkeley Girt English to German automatic run 1) where we used machine translation (L & H Power and the Systran translator) combined with our normalized thesaurus matching technique. This run had better results than the Russian runs, but did not perform comparably to the bilingual main task English-to-German runs.

4 Submissions for the CLEF main tasks

For the CLEF main tasks, we concentrated on French and German as the collection languages and English, French, German and Russian as the topic languages. We participated in 2 tasks: monolingual and bilingual for French and German document collections. We experimented with several translation programs, German decompounding and blind feedback. Two techniques are used almost universally:

Blind Feedback

For our relevance feedback algorithm, we initially searched the collections using the original queries. Then, for each query, we assumed the 20 top-ranked documents to be relevant and selected 30 terms from these documents to add to the original query for a new search.

German decompounding

To decompound the German compounds in the German and Girt collections, we first created a wordlist that included all words in the collections and queries. Using a base dictionary of component words and compounds, we then split the compounds into their components. During indexing, we replaced the German compounds with the component words found in the base dictionary. This technique was first successfully applied by Aitao Chen in the 2001 CLEF competition [6].

4.1 Monolingual Retrieval of the CLEF collections

For CLEF-2002, we submitted monolingual runs for the French and German collections. Our results for the French bilingual runs were slightly better than those for the German runs. In both languages, adding the narrative to the query indexes improved average precision about 6% and 7% for the German and French runs, respectively.

BKMLFF1 (Berkeley Monolingual French against French Automatic Run 1). The original query topics (including title, description and narrative) were searched against the French collection. We applied a blind feedback algorithm for performance improvement. For indexing the French collection, we used a stop word list, the latin-to-lower normalizer and the Muscat French stemmer.

BKMLFF2 (Berkeley Monolingual French against French Automatic Run 2). For indexing and querying the collections, we used the same procedure as in BKMLFF1. For indexing the topics, we only included the title and description.

BKMLGG1 (Berkeley Monolingual German against German Automatic Run 1). The query topics were searched against the German collection. For indexing both the document collection and the queries, we used a stop word list, the latin-to-lower normalizer and the Muscat German stemmer. We used Aitao Chen's decompounding algorithm to split German compounds in both the document collection and the queries. We applied our blind feedback algorithm to the results for performance improvement. All query fields were indexed.

BKMLGG2 (Berkeley Monolingual German against German Automatic Run 2). For this run, we used the same indexing procedure as for BKMLGG1. From the queries, only the title and description were searched against the collections.

4.2 Bilingual Retrieval of the CLEF collections

We submitted 10 bilingual runs for search against the French and German collections. Overall, the Russian to German or French runs yielded decidedly worse results than the other language runs. Submitting English without any translation yielded much worse results than the same experiment in the Amaryllis collection – this was an error in processing where the French stop word list and stemmer were applied to the English topic descriptions instead of the appropriate English ones. Correcting this error (unofficial run BKBIEF2c below) results in an overall precision of 0.2304 instead of the official result of 0.0513.

The English to French runs yielded slightly better results than the English to German runs, whereas the French to German run did better than the German to French run.

4.3 Bilingual to French Documents

Our runs for the CLEF bilingual-to-French main task (as well as monolingual French runs) are summarized in Table 3.

BKBIEF1 (Berkeley Bilingual English against French Automatic Run 1). We translated the English queries with two translation programs: the Systran

Table 3. Results of Berkeley Bilingual to French runs for CLEF-2002.

Run Name	bkmlff1	bkmlff2	bkbief1	bkbief2	bkbief2c	bkbigf1	bkbirf1
Retrieved	50000	50000	50000	50000	50000	50000	50000
Relevant	1383	1383	1383	1383	1383	1383	1383
Rel. ret.	1337	1313	1285	162	874	1303	1211
Precision							
at 0.00	0.8125	0.7475	0.6808	0.0840	0.3137	0.6759	0.5686
at 0.10	0.7747	0.6990	0.6284	0.0795	0.3084	0.6271	0.5117
at 0.20	0.6718	0.6363	0.5642	0.0695	0.2952	0.5582	0.4726
at 0.30	0.5718	0.5358	0.5210	0.0693	0.2817	0.4818	0.4312
at 0.40	0.5461	0.5068	0.4962	0.0672	0.2737	0.4589	0.3841
at 0.50	0.5017	0.4717	0.4702	0.0669	0.2634	0.4389	0.3312
at 0.60	0.4647	0.4332	0.4260	0.0612	0.2462	0.3986	0.3022
at 0.70	0.3938	0.3752	0.3713	0.0481	0.2275	0.3428	0.2656
at 0.80	0.3440	0.3301	0.3302	0.0411	0.1878	0.2972	0.2283
at 0.90	0.2720	0.2666	0.2626	0.0242	0.1401	0.2330	0.1674
at 1.00	0.1945	0.1904	0.1868	0.0188	0.1113	0.1686	0.1093
Avg prec.	0.4884	0.4558	0.4312	0.0513	0.2304	0.4100	0.3276

translator (Altavista Babelfish) and L & H's Power translator. The translations were pooled together and the term frequencies of words occurring twice or more divided (to avoid overemphasis of terms that were translated the same by both programs). The title and description fields of the topics were indexed and searched against the French collections. For indexing the collection, we used the same procedures as in the monolingual runs. For performance improvement, we applied our blind feedback algorithm to the query results.

BKBIEF2 (Berkeley Bilingual English against French Automatic Run 2). We submitted the English queries (all fields) without any translation to the French collections and used the blind feedback algorithm for performance improvement. Collection indexing remained the same.

BKBIGF1 (Berkeley Bilingual German against French Automatic Run 1). We translated the German queries with two translation programs: the Systran translator (Altavista Babelfish) and L & H's Power translator. The translations were pooled together and the term frequencies of words occurring twice or more divided. The title and description fields of the topics were indexed and searched against the French collections. Again, a blind feedback algorithm was applied. Collection indexing remained the same.

BKBIRF1 (Berkeley Bilingual Russian against French Automatic Run 1). We translated the Russian queries with two translation programs: the Systran translator (Altavista Babelfish) and the Promt (http://www.translate.ru/) translator. The Promt translator translated the queries directly from Russian to French, whereas in the Systran translation, we used an intermediate step from the Russian translation to an English translation to than translate further to French (i.e. English is used as a pivot language). The translations were pooled and the title

and description fields submitted to the collection. Our blind feedback algorithm was applied. Collection indexing remained the same.

4.4 Bilingual to German Documents

Our runs for the CLEF bilingual-to-German main task (as well as monolingual German runs) are summarized in Table 4.

Table 4. Results of official Bilingual to German runs for CLEF-2002.

Run name	bkmlgg1	bkmlgg2	bkbifg1	bkbifg2	bkbieg1	bkbieg2	bkbirg1	bkbirg2
Retrieved	50000	50000	50000	50000	50000	50000	50000	50000
Relevant	1938	1938	1938	1938	1938	1938	1938	1938
Rel Ret.	1705	1734	1798	1760	1628	1661	1351	1260
Precision								
at 0.00	0.7686	0.7670	0.8141	0.8122	0.7108	0.6625	0.5638	0.5051
at 0.10	0.6750	0.6161	0.7345	0.6959	0.6190	0.6011	0.5055	0.4029
at 0.20	0.6257	0.5836	0.6959	0.6219	0.5594	0.5595	0.4565	0.3779
at 0.30	0.5654	0.5352	0.5947	0.5565	0.5207	0.5075	0.4141	0.3417
at 0.40	0.5367	0.4983	0.5490	0.5174	0.4741	0.4642	0.3761	0.3202
at 0.50	0.5018	0.4753	0.4851	0.4596	0.4358	0.4359	0.3408	0.2923
at 0.60	0.4722	0.4426	0.4465	0.4226	0.4090	0.4105	0.3122	0.2685
at 0.70	0.4239	0.4027	0.3833	0.3637	0.3647	0.3588	0.2687	0.2375
at 0.80	0.3413	0.3406	0.3084	0.3010	0.2972	0.3061	0.2253	0.1906
at 0.90	0.2642	0.2445	0.2289	0.2191	0.2204	0.2172	0.1659	0.1366
at 1.00	0.1681	0.1451	0.1271	0.1256	0.1441	0.1140	0.0927	0.0720
Bky Avg.	0.4696	0.4404	0.4722	0.4448	0.4150	0.4060	0.3254	0.2691

BKBIEG1 (Berkeley Bilingual English against German Automatic Run 1). We translated the English queries with two translation programs: the Systran translator (Altavista Babelfish) and L & H's Power translator. The translations were pooled together and the term frequencies of words occurring twice or more divided (to avoid overemphasis of terms that were translated the same by both programs). We used the German decompounding procedure to split compounds in the collections and the queries. All query fields were indexed and searched against the German collections. A blind feedback algorithm was applied.

BKBIEG2 (Berkeley Bilingual English against German Automatic Run 1). This resembles BKBIEG1, except that we only submitted the title and description fields of the topics to the German collections.

BKBIFG1 (Berkeley Bilingual French against German Automatic Run 1). We used the same procedures as for the BKBIEG1 run.

BKBIFG2 (Berkeley Bilingual French against German Automatic Run 2). We used the same procedures as for the BKBIEG2 run.

BKBIRG1 (Berkeley Bilingual Russian against German Automatic Run 1). We translated the Russian queries with two translation programs: the Systran

translator (Altavista Babelfish) and the Prompt translator (http://www.translate.ru). The Promt translator translated the queries directly from Russian to German, whereas Systran can only translate from Russian to English. For this Systran translation we then further translated from the English translation into German. Both translation results were pooled and the topics (all fields) submitted to the collection. As before, we used German decompounding for indexing the collections and blind feedback to improve our results.

BKBIRG2 (Berkeley Bilingual Russian against German Automatic Run 2). This resembles BKIRG1, except that we only submitted the title and description fields of the topics to the German collections.

5 Further Analysis of Bilingual Results by Query

After the CLEF workshop we undertook to do some further analysis of the results by query. Of particular interest to us was why our French-to-German Run (BKBIFG1) seemed to have higher average precision (0.4722) than our best German monolingual run (BKMLGG1 which had overall precision of 0.4696). A table of the Bilingual-to-German Runs by Query can be found below as Table 5.

For certain queries, the retrieval performance from a French or Russian language version of the topic (after being translated into German) was much higher than the original German verison of the topic. We postulate three reasons for why the automatic translations of a topic language query to obtain a collection language query might yield better results than the original collection-language version of the topic:

(i) In the machine translation process particular query words might be emphasized through repetition. If those query terms are discriminating, then the ranking will yield better results.

(ii) The translation of the query topics might introduce more specific terms or more general terms (for specific compounds) that help the algorithm in matching more relevant documents.

For example, the translation from French of query 111 had a precision of 0.2807 whereas the monolingual German version only yielded 0.0453. Where the German version of the topic only mentions the compound "Computeranimation" in its phrasing, while the translation from French also introduces the more general terms "Computer" and "Rechner" which found more relevant documents.

(iii) The translation of some query topics might introduce additional important terms or variations of important terms that don't occur in the original query.

For example, topic 139 which asks about EU fishing quotes, had a precision of 0.4059 for the French-German run, where it only yielded 0.0467 for the monolingual German run. Whereas the monolingual query mentions two German compound words "EU-Fischfangquoten" and "Fischereiquoten", the translation from French offers more variety with the phrases "Fangquoten" and "Quoten des

Fischens" and "EU" and "Europäische Union". It is very likely that our decompounding algorithm could not split the two compounds in the monolingual query version, which would mean that any relevant document must have have exactly these two compounds to be retrieved The translated terms, being somewhat less specific, seem to have a much higher chance of matching to relevant German documents.

In another case, topic 131 about intellectual property rights, the German queries perform poorly with the Russian topic (in translation) runs the best performing. This is because the Russian phrase авторских прав (copyright) is translated to the German term Urheberrecht which is missing from the original German formulation of the topic. The Russian-German cross-language performance for this topic is 0.7779 versus the best German monolingual performance of 0.0295.

6 Summary and Acknowledgments

For CLEF-2002, the Berkeley group one concentrated on two collection languages, French and German, and three document collections, Amaryllis, GIRT and CLEF main (French and German newspapers). We worked with four topic languages: English, French, German and Russian. For the three tasks where we worked with Russian as a topic language (GIRT, bilingual Russian to French, and bilingual Russian to German) Russian bilingual consistently underperformed other bilingual topic languages. Why this is the case needs further in-depth investigation. Interestingly enough in the bilingual-to-German documents task, our French topics slightly outperformed our monolingual German runs, retrieving considerably more relevant documents in the top 1000.

Another major focus of our experimentation was to determine the utility of controlled vocabulary and thesauri in cross-language information retrieval. We did experiments with both the Amaryllis and GIRT collections utilizing thesaurus matching techniques. Our results do not show any particular advantage to thesaurus matching over straight translation when machine translation is available; however examination of individual queries shows that thesaurus matching can be a big win sometimes. We are beginning a detailed analysis of individual queries in the CLEF tasks.

This research was supported by research grant number N66001-00-1-8911 (Mar 2000-Feb 2003) from the Defense Advanced Research Projects Agency (DARPA) Translingual Information Detection Extraction and Summarization (TIDES) program. within the DARPA Information Technology Office. We thank Aitao Chen supplying us with his German decompounding software.

References

1. S. Price W. Hersh and L. Donohoe. Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus. In *Proceedings of the 2000 American Medical Informatics Association (AMIA) Symposium*, 2000.

2. A. Chen W. Cooper and F. Gey. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. In D. K. Harman, editor, *The Second Text REtrieval Conference (TREC-2)*, pages 57–66, March 1994.
3. H. Schott (ed.). *Thesaurus for the Social Sciences. [Vol. 1:] German-English. [Vol. 2:] English-German. [Edition] 1999.* InformationsZentrum Sozialwissenschaften Bonn, 2000.
4. M. Kluck and F. Gey. The domain-specific task of clef - specific evaluation strategies incross-language information retrieval. In *Cross Language Retrieval Evaluation, Proceedings of the CLEF 2000 Workshop*, pages 48–56. Springer Computer Science Series LNCS 2069, 2001.
5. H. Jiang F. Gey and N. Perelman. Working with russian queries for the girt, bilingual and multilingual clef tasks. In Carol Peters, Martin Braschler, Julio Gonzales, and Michael Kluck, editors, *Cross Language Retrieval Evaluation, Proceedings of the CLEF 2001 Workshop*, pages 235–243. Springer Computer Science Series LNCS 2406, 2002.
6. A. Chen. Multilingual information retrieval using english and chinese queries. In J. Gonzalo C. Peters, M. Braschler and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001*, pages 44–58. Springer Computer Science Series LNCS 2406, 2001.

Table 5. Bilingual-to-German Precision by Query (* Best performance for topic)

Qry	bkbieg1	bkbieg2	bkbifg1	bkbifg2	bkbirg1	bkbirg2	bkmlgg1	bkmlgg2
91	0.4179	0.3662	0.3936	0.4342*	0.2043	0.0437	0.3080	0.3434
92	0.2481	0.2426	0.2950	0.2700	0.2434	0.2413	0.3299	0.3334
93	0.8737	0.8805	0.8577	0.8806	0.8388	0.0000	0.9098	0.8868
94	0.0040	0.0995	0.1738	0.4649	0.0198	0.3738	0.8850	0.9008
95	0.1624	0.0609	0.1861	0.1893	0.1652	0.1350	0.1490	0.1486
96	0.5341	0.5400	0.6107	0.5772	0.4131	0.4492	0.6141	0.7040
97	0.6633	0.2730	0.7112*	0.3412	0.7095	0.1571	0.8464	0.5706
98	0.6400*	0.6371	0.6184	0.6223	0.5794	0.1927	0.6179	0.6335
99	0.0515	0.2921	0.5770	0.5146	0.0036	0.0113	0.6241	0.4569
100	0.5899	0.6037	0.5416	0.5488	0.5963	0.7601	0.6057	0.5964
101	0.5730	0.7378	0.6448	0.7378	0.4726	0.7216	0.5057	0.7420
102	0.4682	0.4680	0.3962	0.4018	0.4051	0.1371	0.4255	0.4258
103	0.5153	0.5313	0.4106	0.5301	0.2644	0.2290	0.4666	0.5045
104	0.1166	0.1500	0.1553	0.1771	0.0916	0.0471	0.3366	0.2810
105	0.6316	0.6717	0.6110	0.6681	0.5339	0.5826	0.4665	0.2042
106	0.0496	0.0907	0.1156	0.0730	0.0968	0.1290	0.1309	0.2375
107	0.0350	0.0959	0.0952	0.1338	0.0206	0.0186	0.0719	0.1126
108	0.3834	0.3545	0.4146	0.4459*	0.0000	0.0000	0.4175	0.4233
109	0.2370	0.1352	0.4032	0.0202	0.3397	0.0056	0.1793	0.0006
110	0.6333	0.5794	0.6351	0.6194	0.6614	0.6535	0.5800	0.5573
111	0.2629	0.0620	0.2526	0.2807*	0.0726	0.0010	0.0825	0.0453
112	0.6317	0.6203	0.7129	0.6168	0.5901	0.6051	0.6229	0.6194
113	0.2819*	0.0020	0.2167	0.0279	0.0002	0.0001	0.1216	0.2623
114	0.4456	0.4315	0.4932	0.4330	0.5178	0.5273*	0.4547	0.4261
115	0.3325	0.2794	0.4024	0.4120	0.3109	0.2368	0.3580	0.2034
116	0.7387	0.7100	0.7497	0.6823	0.0169	0.0000	0.7008	0.7067
117	0.0000	0.0000	0.0000	0.0000	0.3035	0.3834*	0.5523	0.6375
118	0.0065	0.0095	0.2845	0.0870	0.0059	0.0422	0.4267	0.1233
119	0.8539	0.8770	0.8368	0.8477	0.8707	0.8734	0.8477	0.8198
120	0.0444	0.0274	0.0500	0.0355	0.0188	0.0494	0.0183	0.0426
121	0.2535	0.4556	0.3500	0.3687	0.1682	0.0387	0.4514	0.3833
122	0.0306	0.0804	0.0972	0.1992	0.0195	0.0557	0.0389	0.0241
123	0.9571	0.8833	0.8755	0.8929	1.0000*	0.9571	0.8833	0.8929
124	0.2980	0.2875	0.5175*	0.4557	0.0202	0.0279	0.5805	0.4533
125	0.5580	0.5283	0.5445	0.5511	0.6788*	0.6069	0.6196	0.5425
126	0.0479	0.0435	0.4758	0.0438	0.4062	0.0435	0.2315	0.0437
127	0.8892	0.8157	0.8864	0.8913	0.0077	0.0010	0.8973	0.8996
128	0.1193	0.0488	0.2852*	0.1776	0.0076	0.0092	0.0372	0.1307
129	0.6476	0.7195*	0.7183	0.6855	0.5968	0.7132	0.6964	0.6890
130	0.6743	0.5810	0.6740	0.5824	0.5862	0.5686	0.8487	0.7301
131	0.4329	0.5522	0.5366	0.5514	0.7779*	0.6080	0.0295	0.0114
132	0.7223	0.7353	0.4720	0.5506	0.7904	0.7724	0.6380	0.7202
133	0.5709	0.4913	0.6386*	0.5993	0.0210	0.0004	0.7521	0.7082
134	0.6060	0.5667	0.6573	0.5612	0.2952	0.0632	0.5802	0.6619
135	0.0762	0.2821	0.3388	0.2472	0.4585	0.2393	0.6250	0.3091
136	1.0000	0.9683	1.0000	1.0000	0.0000	0.0000	1.0000	0.9481
137	0.0164	0.0417	0.0019	0.0000	0.0175	0.0256	0.0000	0.0000
138	0.8533	0.8514	0.7679	0.7024	0.8404	0.8607*	0.8772	0.8052
139	0.0102	0.0685	0.3312	0.4059*	0.2110	0.2438	0.0386	0.0467
140	0.0002	0.0005	0.0716	0.0516	0.0020	0.0148	0.0004	0.0718
Avg.	0.4150	0.4060	0.4722	0.4448	0.3254	0.2691	0.4696	0.4404